Supplemental Data

American Journal of Human Genetics, Volume *88*

DASH: A Method for Identical by Descent-Based Haplotype Discovery Uncovers Association to Recent Variation

Alexander Gusev, Eimear E. Kenny, Jennifer K. Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M. Altshuler, Jeffrey M. Friedman, Jan L. Breslow, and Itsik Pe'er

Table S1. Data Set Summary

	Kosrae	WTCCC
Total samples	2,906	16,179
Total polymorphic variants	398,876	451,490
Hidden variants	121,633	93,896
in reference panel	98,488	75,845
Simulated causal variants	12,500	12,500
in reference panel	10,380	10,333
Total tested variants	277,243	357,594
Genome-wide significance threshold (DASH)	4.0×10^{-8}	5.4×10^{-8}
Genome-wide significance threshold (GWAS)	2.5×10^{-7}	2.0×10^{-7}
Genome-wide significance threshold (imputation)	2.1×10^{-7}	9.3×10^{-8}

Table S2. All Genome-Wide Significant Regions Identified in the WTCCC

Trait	Chr	Haplotype start	Haplotype end	*f*	OR	*P* DASH[1]	Controls only *P* GWAS[2]	Pooled controls *P* GWAS[3]	conditioned *P* DASH[4]	stepwise conditioned *P* DASH[5]	Published relevant associations	r^2 in 0.95 calls[6]	r^2 in 0.98 calls[7]
CAD	6	160,730,710	160,927,528	1.3%	2.13	8.2E-11	1.4E-04	5.9E-05	2.3E-06	2.3E-10	*SLC22A3,LPAL2,LPA*	0.98	0.97
CAD	6	160,927,528	160,930,756	1.2%	2.05	3.4E-09	1.4E-04	5.9E-05	1.4E-06	7.1E-09	*SLC22A3,LPAL2,LPA*	0.64	0.64
CAD	9	22,062,730	22,227,321	34.9%	0.78	1.4E-11	7.7E-15	2.9E-17	*NS*	*NS*	*CDKN2A*	0.90	0.90
CD	1	67,374,727	67,523,062	14.8%	1.38	5.1E-12	5.0E-19	1.2E-24	*NS*	*NS*	*IL23R*	0.94	0.88
CD	5	131,829,846	131,844,802	37.7%	1.29	4.2E-12	8.1E-08	8.2E-12	*NS*	*NS*	5q31 region	0.89	0.88
CD	16	49,152,847	49,164,501	3.7%	1.95	9.9E-19	9.9E-16	7.5E-19	1.4E-08	1.4E-08	*NKD1*	0.89	0.28
CD	16	49,477,102	49,477,284	1.6%	2.10	5.3E-12	9.9E-16	7.5E-19	5.1E-07	5.1E-07	*NOD2*	0.96	0.95
RA	6	33,903,649	33,907,610	1.1%	2.18	1.4E-09	2.7E-42	9.4E-64	4.1E-07	1.2E-05	MHC	0.64	0.11
T1D	6	28,110,942	28,119,631	0.8%	3.57	1.5E-23	4.2E-144	4.9E-175	2.1E-04	1.5E-04	MHC	0.99	0.99
T1D	6	28,476,919	28,478,225	12.1%	1.61	4.1E-25	4.2E-144	4.9E-175	2.1E-02	1.4E-02	MHC	0.77	0.77
T1D	**6**	**33,787,785**	**33,790,392**	**0.6%**	**3.94**	**4.2E-24**	**4.2E-144**	**4.9E-175**	**2.7E-13**	**1.0E-07**	**MHC**	**0.99**	**0.99**

Haplotypes also shown in Table 1 are marked in bold

[1]Strongest association in locus, conditionally independent of all genome-wide significant and local haplotypes on chromosome

[2]Most significant nearby single marker association with NBS & 58C controls only

[3]Most significant nearby single marker association with combined alternative cases and controls

[4]Least significant haplotype association after conditioning on any nearby single markers (combined cases and controls)

[5]Haplotype association after conditioning on all independently genome-wide significant single markers (combined case and controls)

[6]Strongest correlation (r^2) to a haplotype identified in a subset of markers with at least 98% of samples having genotype call confidence of 0.95

[7]Strongest correlation (r^2) to a haplotype identified in a subset of markers with at least 98% of samples having genotype call confidence of 0.98

Table S3. All Genome-Wide Significant, Localized Regions Identified in the WTCCC

Trait	Chr	Haplotype start	Haplotype end	*f*	OR	*P* DASH[1]	Controls only *P* GWAS[2]	Pooled controls *P* GWAS[3]	conditioned *P* DASH[4]	stepwise conditioned *P* DASH[5]	Published relevant associations	r^2 in 0.95 calls[6]	r^2 in 0.98 calls[7]
CAD	**6**	**160,932,260**	**161,017,268**	**1.7%**	**2.17**	**4.1E-14**	**1.4E-04**	**5.9E-05**	**2.6E-09**	**4.1E-14**	***SLC22A3,LPAL2,LPA***	**1.00**	**1.00**
CAD	9	22,021,005	22,109,128	35.8%	1.30	4.1E-14	7.7E-15	2.9E-17	*NS*	*NS*	*CDKN2A*	1.00	0.99
CD	1	67,283,445	67,348,663	4.5%	0.39	8.0E-15	5.0E-19	1.2E-24	5.7E-04	1.2E-03	*IL23R*	0.99	0.99
CD	1	67,424,321	67,478,314	25.3%	1.43	1.9E-20	5.0E-19	1.2E-24	*NS*	*NS*	*IL23R*	0.99	0.99
CD	5	131,703,880	131,782,903	42.2%	1.29	6.8E-13	8.1E-08	8.2E-12	4.6E-02	4.6E-02	5q31 region	0.97	0.93
CD	**16**	**49,244,058**	**49,244,516**	**7.2%**	**1.80**	**1.7E-24**	**9.9E-16**	**7.5E-19**	**3.0E-10**	**3.0E-10**	***NKD1***	**0.98**	**0.98**
RA	**6**	**33,512,042**	**33,684,663**	**2.4%**	**2.35**	**1.0E-23**	**2.7E-42**	**9.4E-64**	**1.8E-16**	**3.3E-13**	***MHC***	**1.00**	**1.00**
T1D	6	26,507,565	26,509,417	12.4%	1.43	4.0E-14	4.2E-144	4.9E-175	*NS*	1.6E-02	*MHC*	0.99	0.99
T1D	6	26,622,750	26,636,229	0.6%	3.78	1.1E-21	4.2E-144	4.9E-175	1.6E-03	1.8E-05	*MHC*	1.00	1.00
T2D	**11**	**22,429,718**	**22,483,326**	**0.3%**	**3.79**	**1.9E-10**	**2.8E-03**	**3.1E-03**	**4.3E-08**	**1.9E-10**	**-**	**0.99**	**0.99**

**Haplotypes also shown in Table 1 are marked in bold*

[1]Strongest association in locus, conditionally independent of all genome-wide significant and local haplotypes on chromosome

[2]Most significant nearby single marker association with NBS & 58C controls only

[3]Most significant nearby single marker association with combined alternative cases and controls

[4]Least significant haplotype association after conditioning on any nearby single markers (combined cases and controls)

[5]Haplotype association after conditioning on all independently genome-wide significant single markers (combined case and controls)

[6]Strongest correlation (r^2) to a haplotype identified in a subset of markers with at least 98% of samples having genotype call confidence of 0.95

[7]Strongest correlation (r^2) to a haplotype identified in a subset of markers with at least 98% of samples having genotype call confidence of 0.98

Table S4. All Genome-Wide Significant Regions Identified in Kosraen Data

Trait	Chr	Haplotype start	Haplotype end	*f*	Hom ref[1]	Het[1]	Hom mut[1]	OR	Sequenced	*P* DASH[2]	*P* GWAS[3]	conditioned *P* DASH[4]	Published relevant associations
Folate	**19**	**10,738,387**	**10,755,240**	**3.2%**	**1,758**	**118**	**6**	**1.68**	**1**	**6.4E-08**	**6.7E-05**	**5.1E-05**	***LDLR; TYK2***
HBA1C	16	87,308,019	87,339,751	4.4%	981	98	0	0.52	2	2.0E-09	6.3E-08	4.3E-05	
HBA1C	16	88,656,265	88,690,776	2.9%	1,032	45	2	0.45	1	8.6E-09	9.4E-04	1.0E-06	
LDL	**12**	**99,350,824**	**99,608,800**	**2.8%**	**2,615**	**160**	**0**	**0.59**	**0**	**1.0E-08**	**1.0E-04**	**2.5E-06**	
Total Cholesterol	**6**	**161,091,801**	**161,107,669**	**13.4%**	**2,065**	**642**	**46**	**1.33**	**1**	**6.9E-11**	**2.5E-06**	**4.9E-06**	***LPA***
Total Cholesterol	**12**	**99,321,530**	**99,350,824**	**2.7%**	**2,603**	**150**	**0**	**0.54**	**0**	**9.1E-11**	**5.9E-04**	**1.6E-08**	
Triglycerides	11	116,105,036	116,272,109	32.1%	1,245	1,236	273	1.27	4	2.6E-13	3.0E-12	0.01	*APOA1, APOA5*
TSH	9	97,467,378	97,674,793	14.0%	1,371	429	58	0.71	1	2.6E-11	2.4E-14	*NS*	
Uric Acid	**11**	**64,653,308**	**64,661,484**	**1.8%**	**1,796**	**65**	**0**	**0.13**	**2**	**5.5E-48**	**3.7E-17**	**2.3E-33**	***SLC22A11-12***
Uric Acid	**19**	**63,081,167**	**63,093,854**	**1.6%**	**1,797**	**64**	**0**	**0.46**	**0**	**9.3E-08**	**3.4E-03**	**9.5E-06**	

[1]Number of phenotyped carriers
[2]Most significant association in locus, conditionally independent of all nearby haplotypes
[3]Most significant nearby single marker association
[4]Least significant haplotype association after conditioning on all nearby single markers

Table S5. All Genome-Wide Significant, Localized Regions Identified in Kosraen Data

Trait	Chr	Haplotype start	Haplotype end	*f*	Hom ref[1]	Het[1]	Hom mut[1]	OR	Sequenced	*P* DASH[2]	*P* GWAS[3]	conditioned *P* DASH[4]	Published relevant associations
C-Reactive Protein	1	156,427,059	156,492,709	24.7%	1,048	713	111	1.28	3	8.2E-09	9.0E-09	*NS*	*SPTA1*
C-Reactive Protein	19	50,153,836	50,169,221	12.3%	1,413	434	25	0.73	1	3.6E-09	6.5E-13	*NS*	*APOE, APOC*
HBA1C	**16**	**87,261,769**	**87,465,349**	**9.9%**	**872**	**201**	**6**	**0.47**	**3**	**2.5E-24**	**6.3E-08**	**2.1E-17**	
LDL	**19**	**46,645,270**	**46,787,809**	**1.0%**	**2,719**	**56**	**0**	**0.45**	**0**	**2.1E-08**	**1.7E-06**	**1.0E-03**	
LDL	19	50,114,427	50,255,095	12.4%	2,120	621	34	1.43	1	5.0E-17	1.1E-23	*NS*	*APOE, APOC*
Total Cholesterol	**11**	**116,106,921**	**116,111,104**	**23.5%**	**1,587**	**1,024**	**142**	**1.21**	**4**	**8.5E-08**	**2.0E-05**	**9.1E-04**	***APOA1, APOA5***
Total Cholesterol	19	50,065,116	50,254,711	12.4%	2,103	616	34	1.40	1	5.8E-15	8.7E-18	*NS*	*APOE, APOC*
Triglycerides	**11**	**116,147,610**	**116,157,417**	**27.4%**	**1,446**	**1,103**	**205**	**1.31**	**4**	**2.2E-15**	**3.0E-12**	**6.8E-05**	***APOA1, APOA5***
TSH	9	97,613,702	97,676,546	18.1%	1,261	507	90	0.72	1	1.6E-12	2.4E-14	*NS*	*PTCH1*
Uric Acid	11	64,117,281	64,341,535	1.8%	1,796	65	0	0.13	2	6.6E-49	9.6E-35	2.9E-17	*SLC22A11-12*

[1]Number of phenotyped carriers
[2]Most significant association in locus, conditionally independent of all nearby haplotypes
[3]Most significant nearby single marker association
[4]Least significant haplotype association after conditioning on all nearby single markers

Table S6. Phenotype Summary for Kosraen Data

Trait	Abbreviation	Total Samples
Body Mass Index	BMI	2,369
Cornell voltage	CLVresstd	1,576
C-Reactive Protein	logCRP	1,872
Diastolic Blood Pressure	DBP	2,520
Fasting Blood Sugar	FBS	1,753
Folic Acid	Folate	1,882
Height	HeightZ	2,364
Hemoglobin A1c	HBA1C	1,079
High Density Lipoprotein Cholesterol	HDL	2,774
Homocysteine	Homocysteine	1,870
Insulin Sensitivity	INSD	1,166
Leptin	LEP	2,849
Low Density Lipoprotein Cholesterol	LDL	2,775
Percentage of Body Fat	pctfat	1,796
PR interval	PRresstd	1,595
QRS interval	QRSdurresstd	1,573
Resting rate interval	RRresstd	1,513
Sokolow-Lyon voltage	SLVresstd	1,541
Systolic Blood Pressure	SBP	2,521
Thyroid Stimulating Hormone	TSH	1,858
Total Cholesterol Levels	TC	2,753
Triglyceride	TG	2,754
Uric Acid	logUrate	1,861
Waist Circumference	waist	2,814
Weight	WT	2,335

Table S7. Phenotype Summary for WTCCC Data

Trait		Cases
Bipolar Disorder	BD	1,998
Coronary Artery Disease	CAD	1,988
Chron's Disease	CD	2,005
Hypertension	HT	2,001
Rheumatoid Arthritis	RA	1,999
Type 1 Diabetes	T1D	2,000
Type 2 Diabetes	T2D	1,999

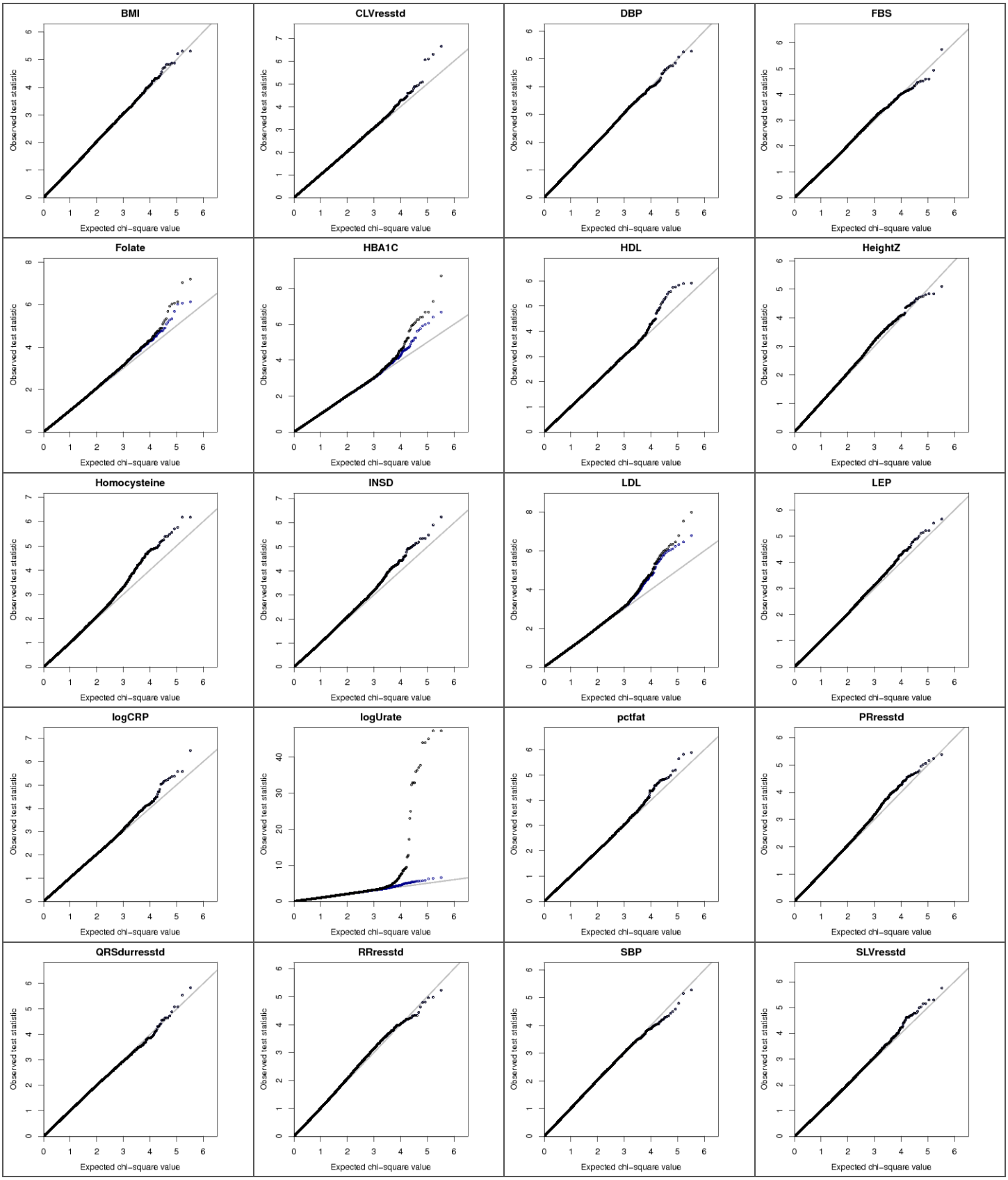
Table S8. Initial and Replicated Haplotype Association with HBa1c

Cohort	Start	End	SNPs	*f*[1]	PVAL	corrected[2] PVAL	OR
Kosrae	87,261,769	87,465,349	10	9.95%	2.5E-24	8.5E-19	0.47
DGI replication	87,404,625	87,560,132	16	0.63%	3.9E-05	0.015	0.19

[1]Frequency in entire cohort
[2]Bonferroni corrected for all haplotypes tested

Table S9. Functional Variants Present in Carriers of Two Associated Haplotypes

Haplotype carriers					Haplotype non-carriers										
avg cov	no call	wild-type	het.	hom. var	avg cov	no call	wild-type	het.	hom. var	chr	Position	Gene	dbSNP	SiFT prediction	Haplotype-SNV correlation
2	2	0	1	0	1	1	3	0	0	16	87,240,737	CYBA	rs4673:G	TOLERATED	0.003
2	2	0	0	1	1	4	0	0	0	16	87,306,116	C16orf84	rs7205989:A	DAMAGING	0.051
1	2	0	0	1	0	4	0	0	0	16	87,399,646	CDT1	rs507329:C	TOLERATED	-
5	**0**	**0**	**2**	**1**	**1**	**3**	**1**	**0**	**0**	**16**	**87,821,231**	**ZNF778**	**novel**	**DAMAGING**	**0.001**
2	2	0	0	1	2	3	1	0	0	16	87,875,539	ANKRD11	novel	TOLERATED	0.122
12	0	2	1	0	9	0	4	0	0	16	87,877,261	ANKRD11	novel	TOLERATED	-
8	**0**	**1**	**2**	**0**	**7**	**0**	**4**	**0**	**0**	**16**	**87,878,113**	**ANKRD11**	**novel**	**TOLERATED**	**-**


BMI
CLVresstd
DBP
FBS
Folate
HBA1C
HDL
HeightZ
Homocysteine
INSD
LDL
LEP
logCRP
logUrate
pctfat
PRresstd
QRSdurresstd
RRresstd
SBP
SLVresstd
Observed test statistic
Expected chi-square value

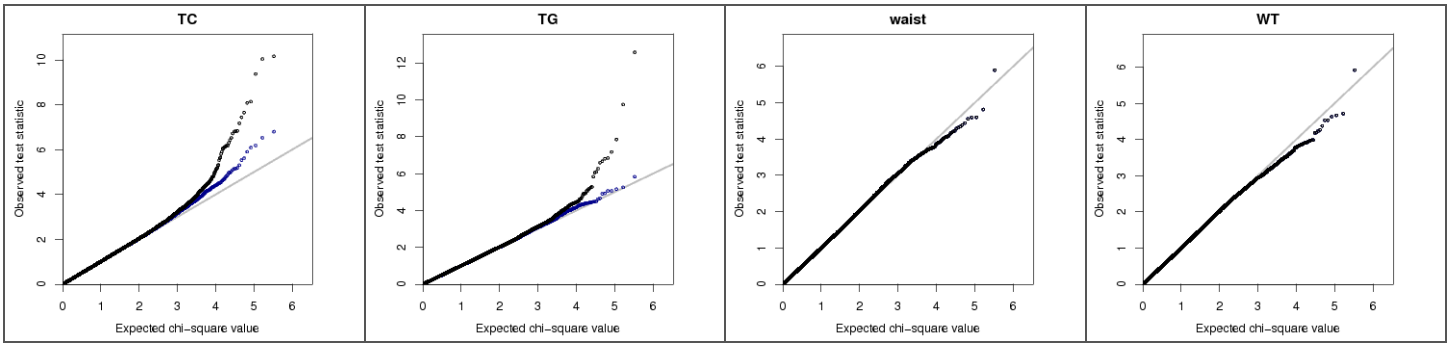


Figure S1. Quantile-Quantile Plot of All Phenotypes Analyzed in Kosraen Cohort

For each of the seven disease traits, a quantil-quantile plot of the allelic association test across all DASH haplotypes having carrier frequency >0.1% is shown in black. Quantile-quantile plot after removing all haplotypes in regions of significant association (Table 1, Supplementary Table 2) is overlaid in blue. Grey line shows the null distribution.

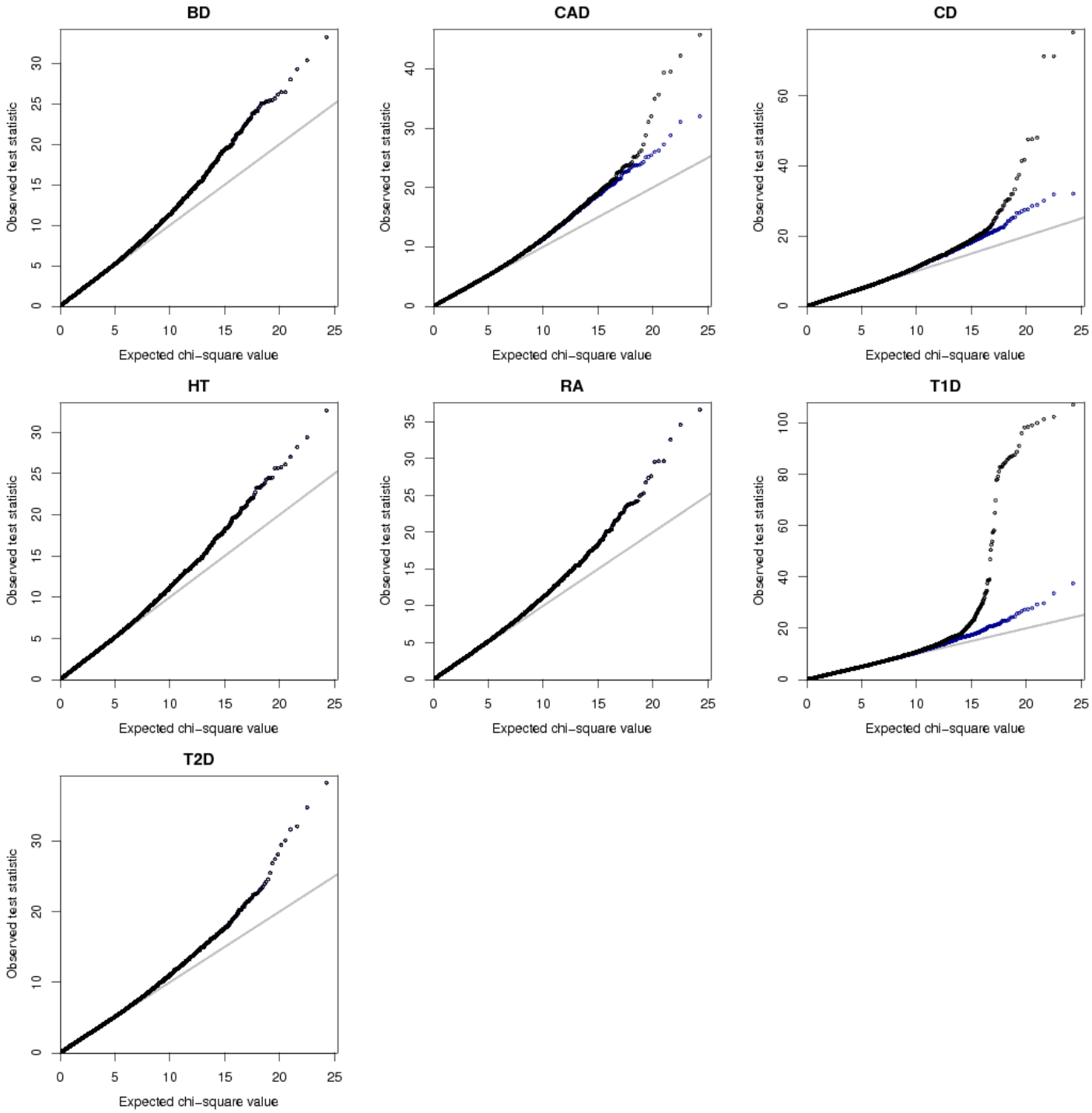
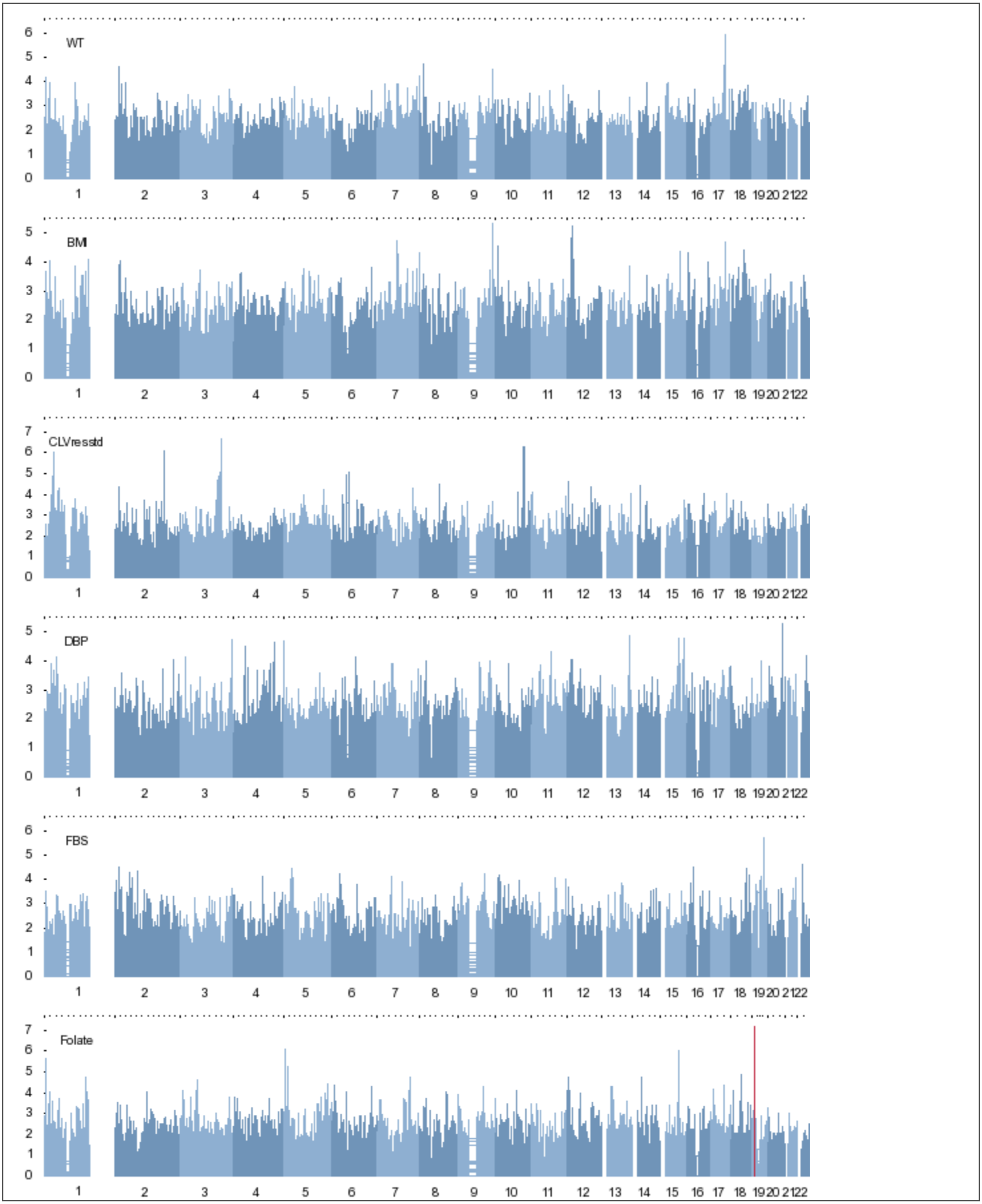
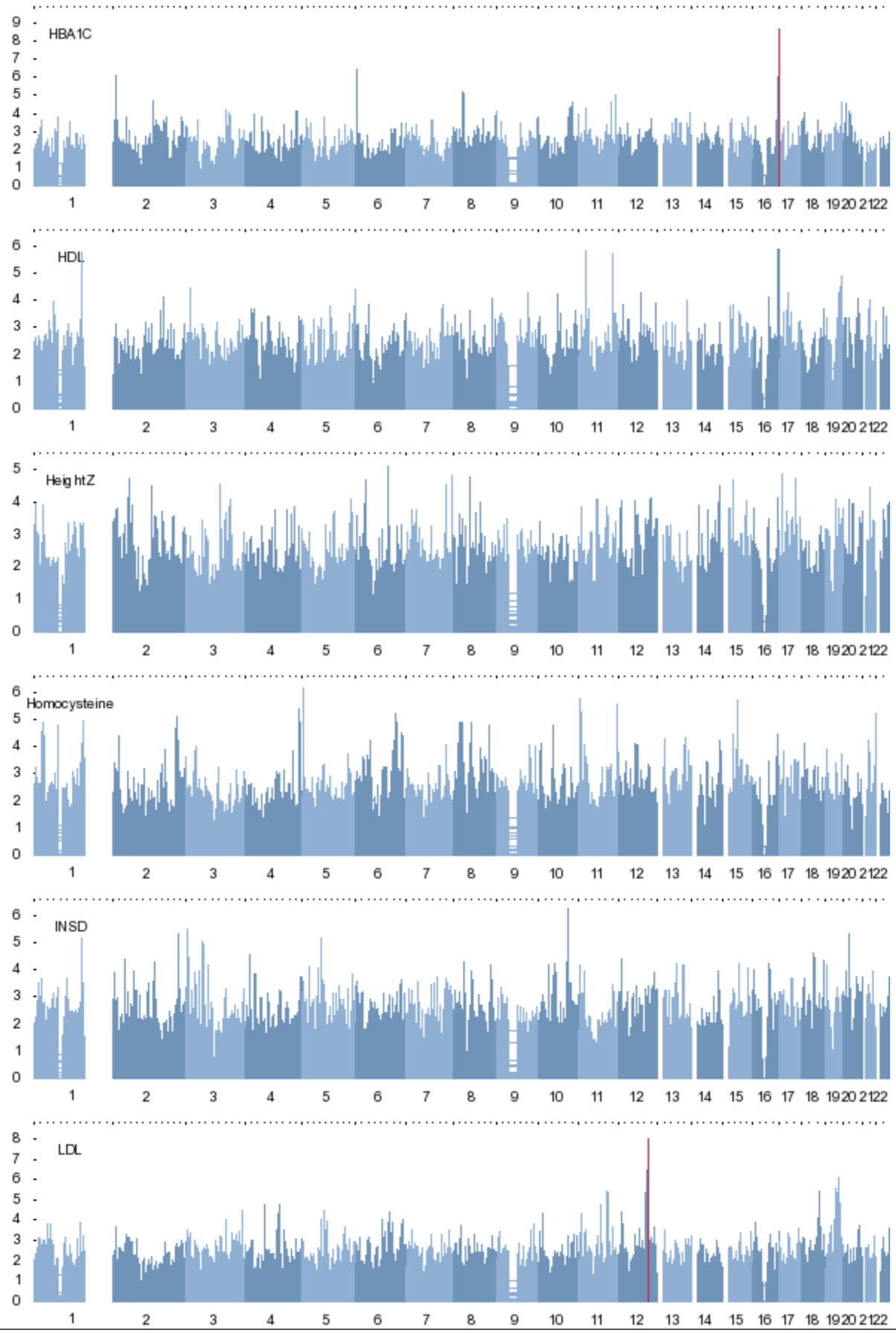
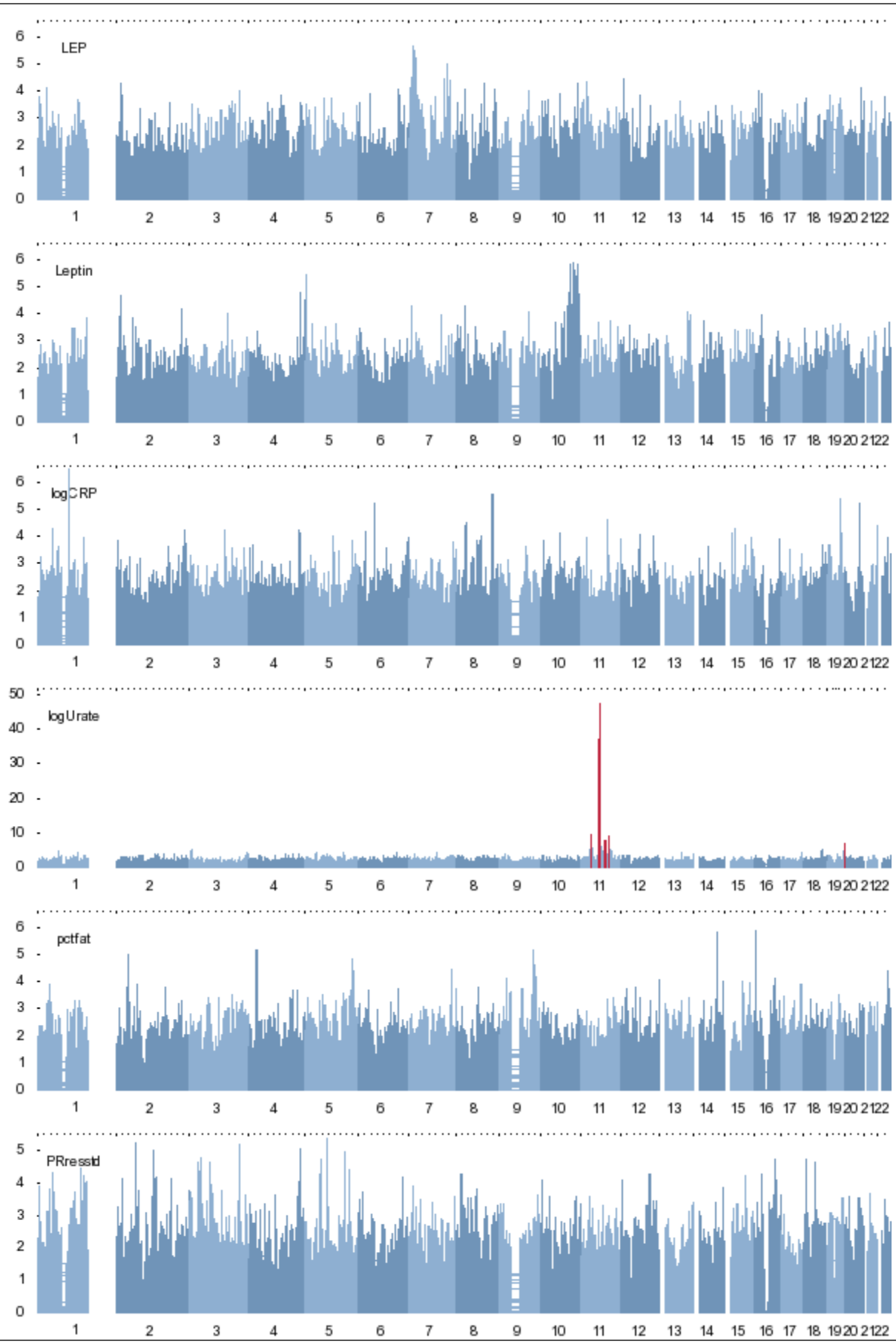


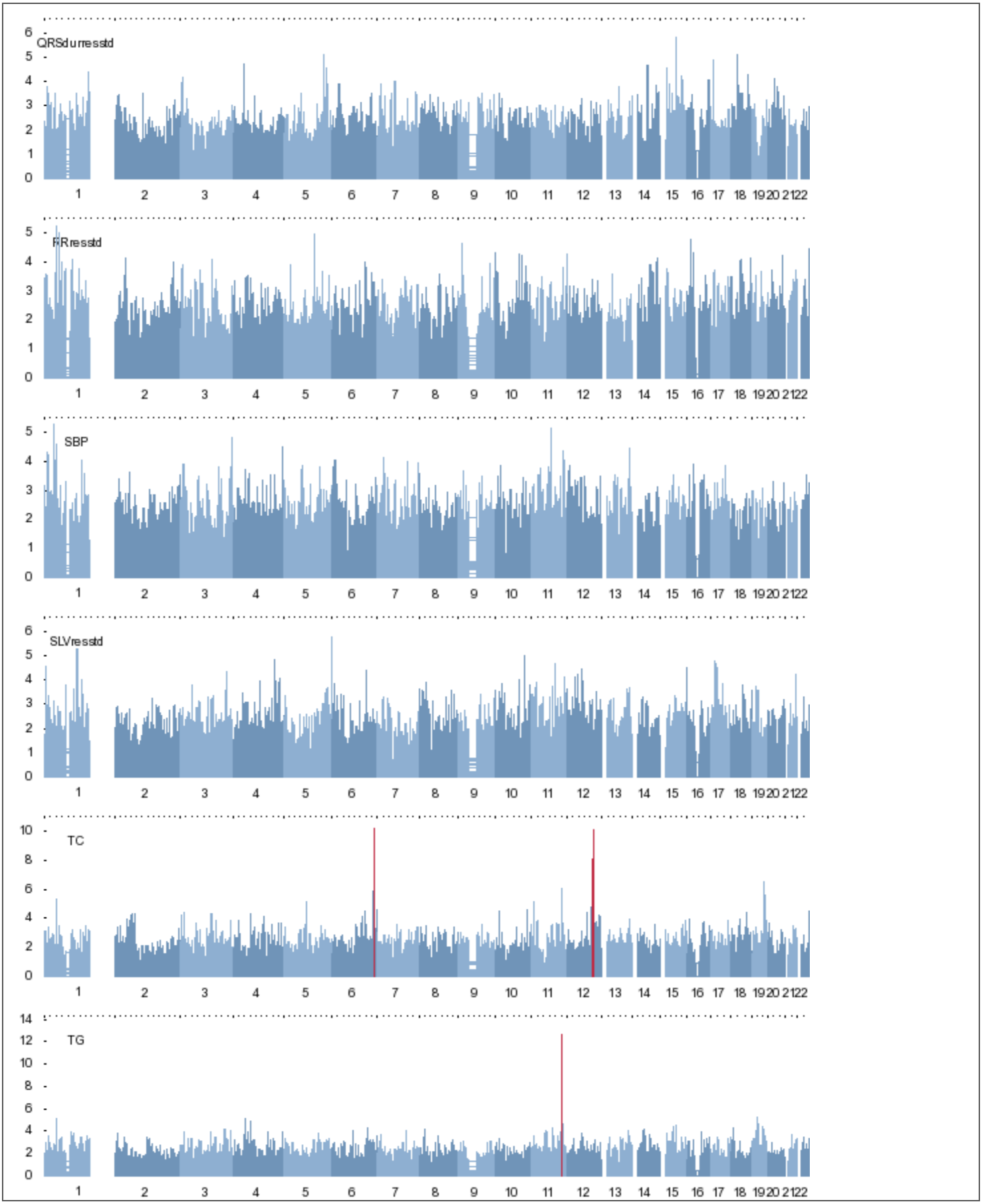
Figure S2. Quantile-Quantile Plot of all phenotypes Analyzed in WTCCC Cohort

For each of the seven disease traits, a quantil-quantile plot of the allelic association test across all DASH haplotypes having carrier frequency >0.1% is shown in black. Quantile-quantile plot after removing all haplotypes in regions of significant association (Table 1, Supplementary Table 2) is overlaid in blue. Grey line shows the null distribution.


WT
BMI
CLVresstd
DBP
FBS
Folate
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


HBA1C
HDL
HeightZ
Homocysteine
INSD
LDL
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


LEP
Leptin
logCRP
logUrate
pctfat
PRresstd
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


QRSdurresstd
RRresstd
SBP
SLVresstd
TC
TG
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

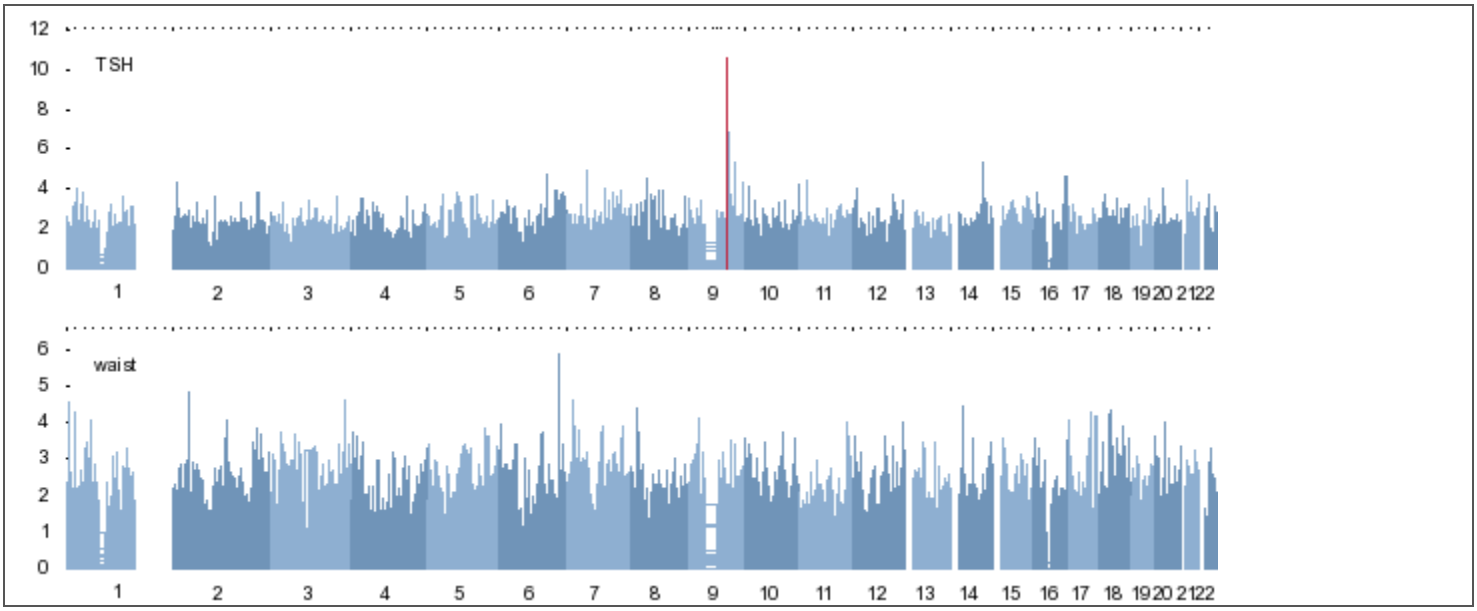
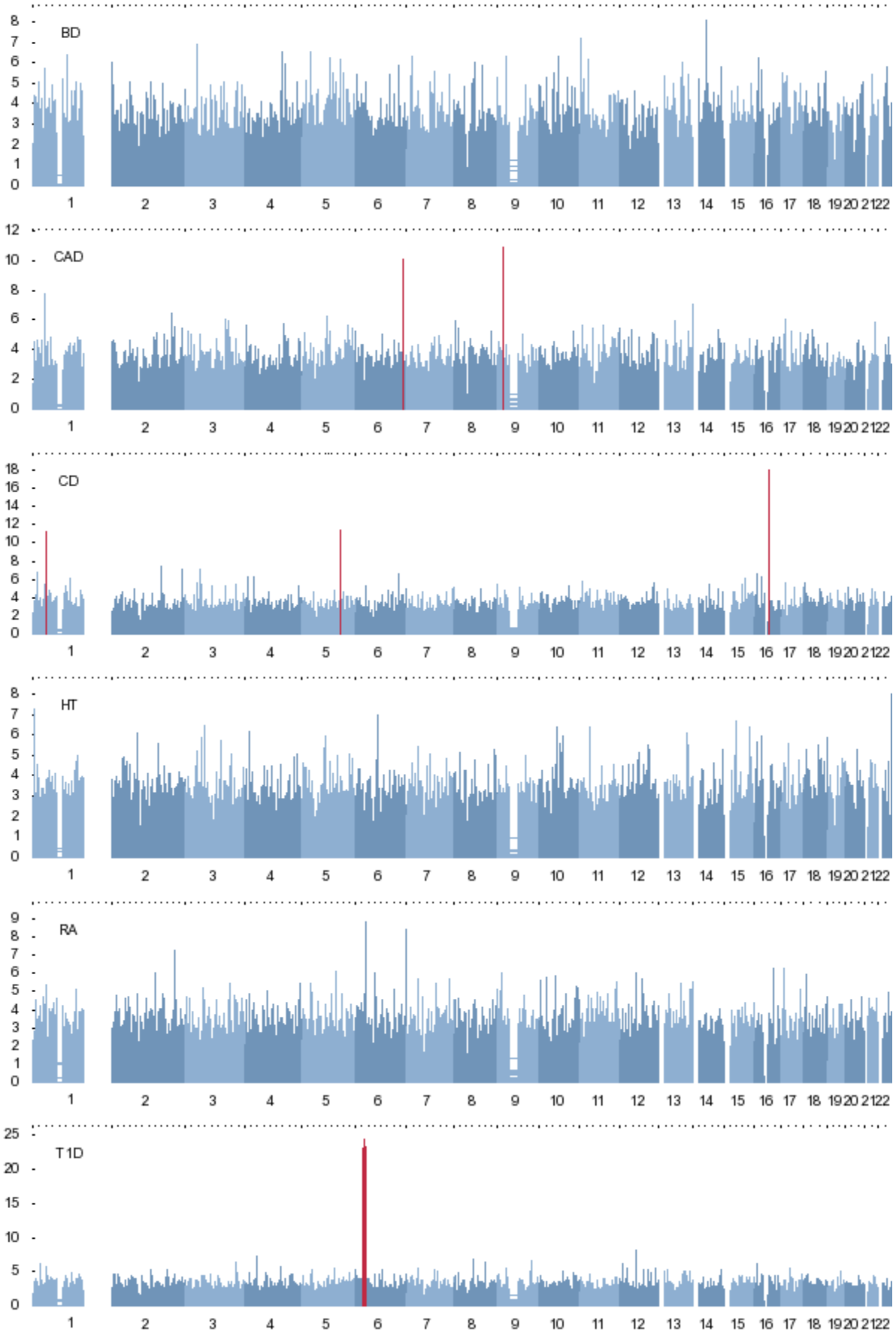


Figure S3. Manhattan Haplotype Plot of All Phenotypes Analyzed in Kosraen Cohort

For each of the seven disease traits -log10 of the allelic P-value for all DASH haplotypes having carrier frequency >0.1% are plotted as bars corresponding to their position in the chromosome. Chromosomes are shown in alternating colors with haplotypes surpassing genome-wide significance ($1.5x10^{-7}$) shown in red.


BD
CAD
CD
HT
RA
T1D
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

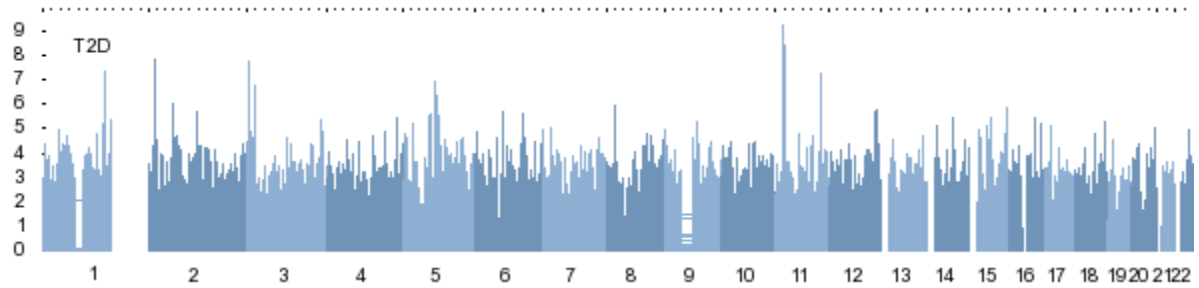
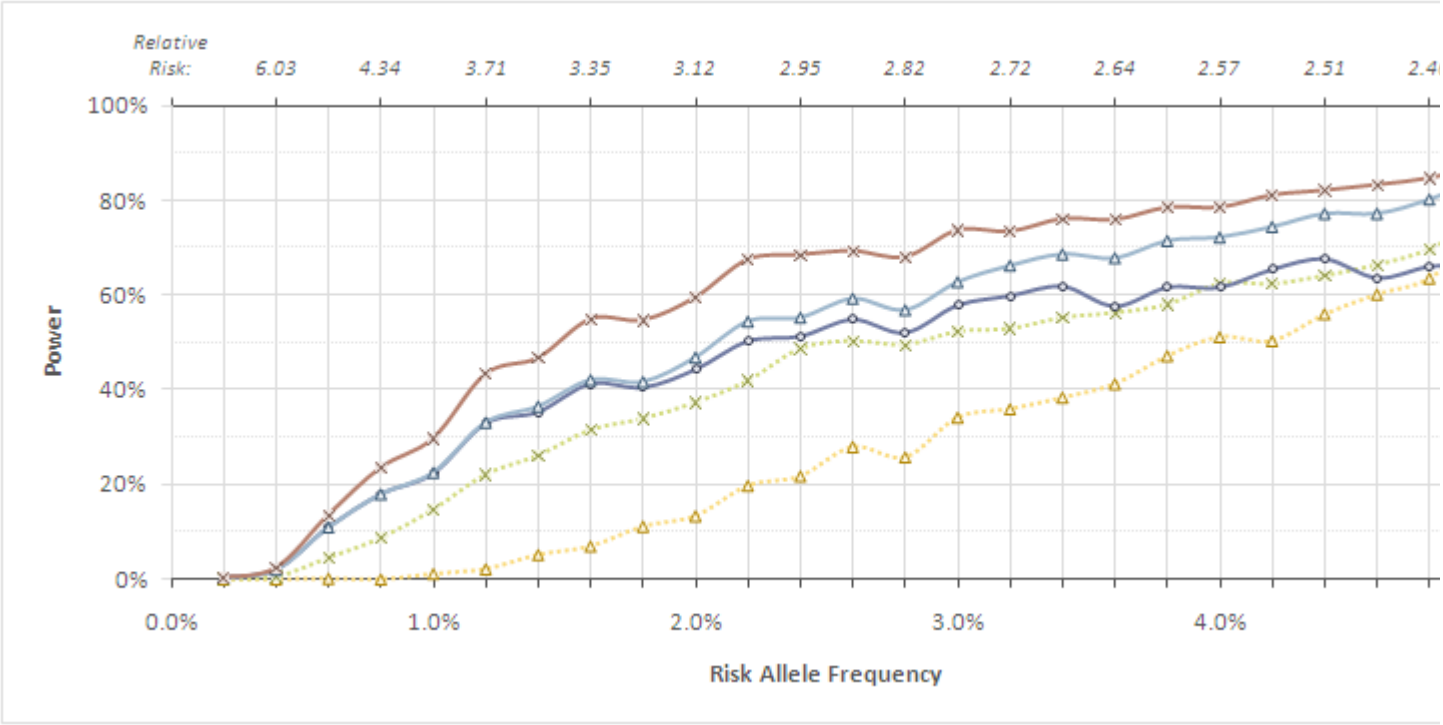


Figure S4. Manhattan Haplotype Plot of All Phenotypes Analyzed in WTCCC Cohort

For each of the seven disease traits -log10 of the allelic P-value for all DASH haplotypes having carrier frequency >0.1% are plotted as bars corresponding to their position in the chromosome. Chromosomes are shown in alternating colors with haplotypes surpassing empirical genome-wide significance (2.5×10^{-8}) shown in red.

A: Association power in Kosraen cohort



B: Association power in WTCCC

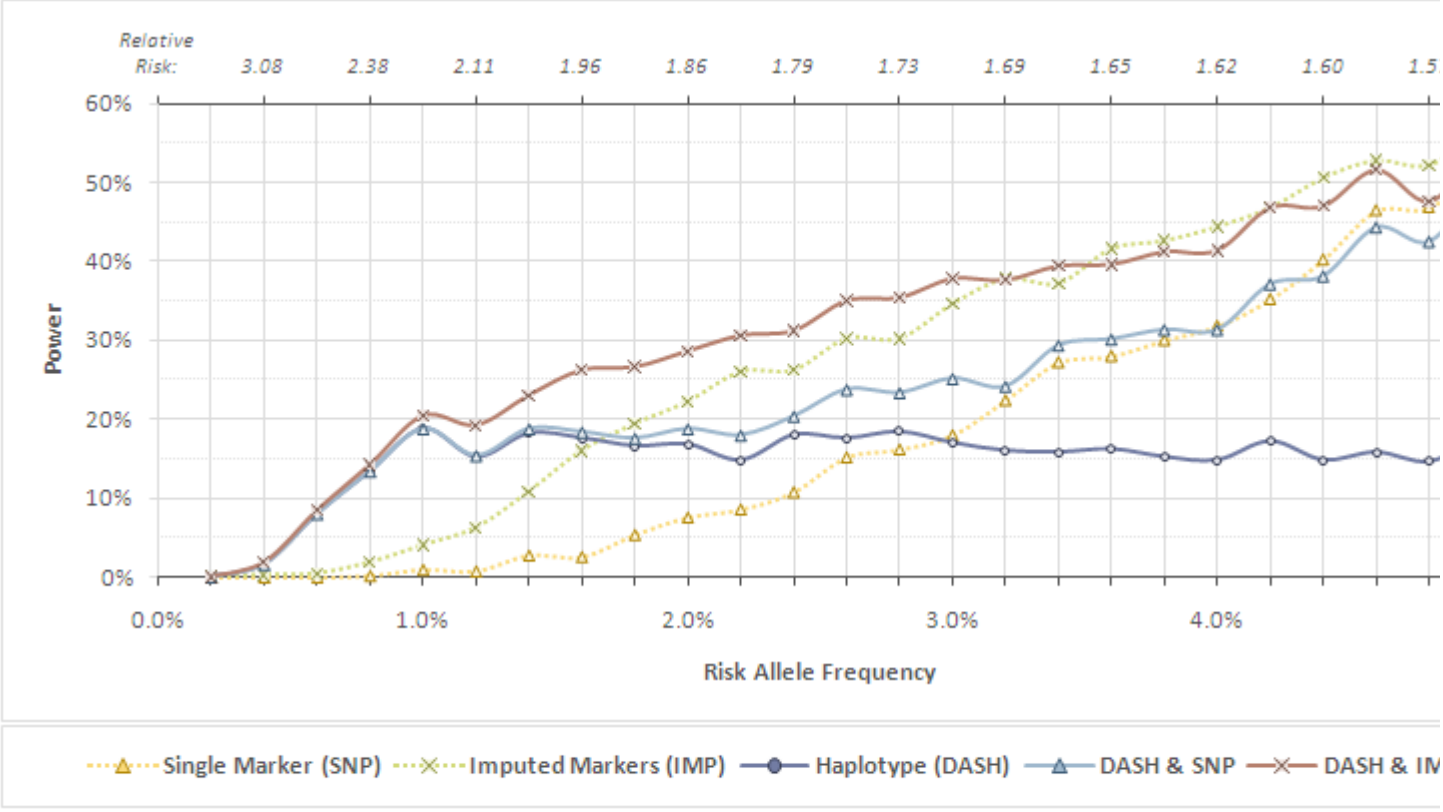


Figure S5. Detailed Simulated Power Analysis

Power to detect a single rare variant at risk allele frequency range of (0%,5%) with proxy markers of MAF ≥ 5%. Lower x-axis details planted variant allele frequency and upper x-axis details corresponding relative risk. Tested separately were single markers (yellow triangle, SNP); imputation from HapMap reference and single markers (green cross, IMP); DASH

haplotypes (blue circle, DASH); DASH haplotypes and single markers (blue triangle, DASH & SNP); DASH haplotypes and imputed markers (brown cross, DASH & IMP). Each point represents power to detect genome-wide significant association at 500 randomly planted markers. Panel A plots results in isolate cohort from Kosrae, Micronesia (imputed from JPTCHB reference); Panel B plots results in European cohort from WTCCC (imputed from CEU reference).

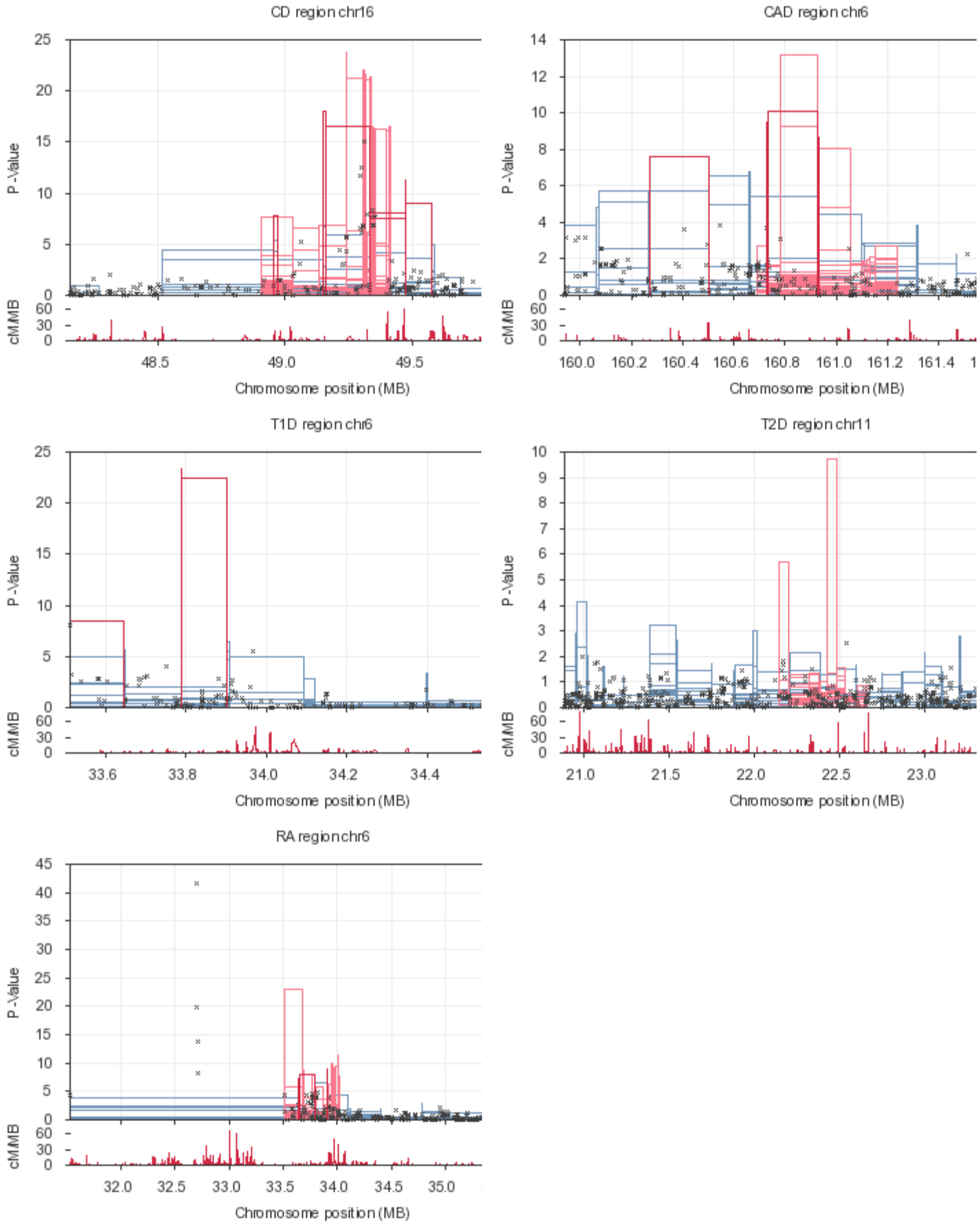
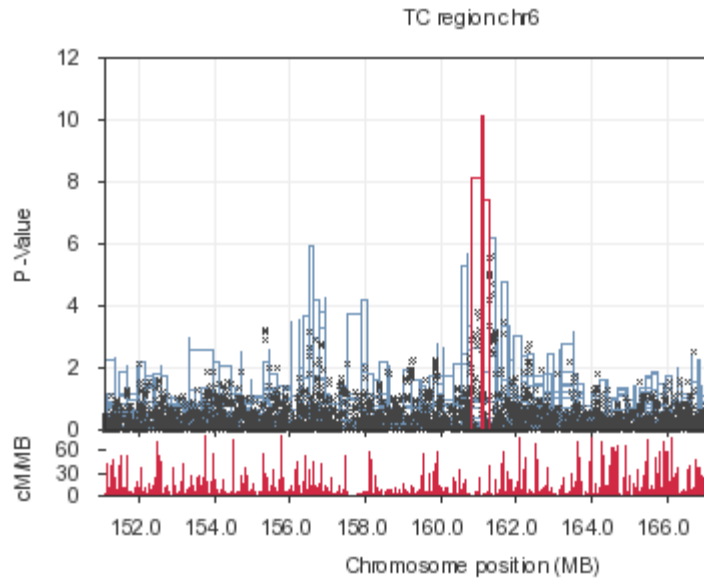
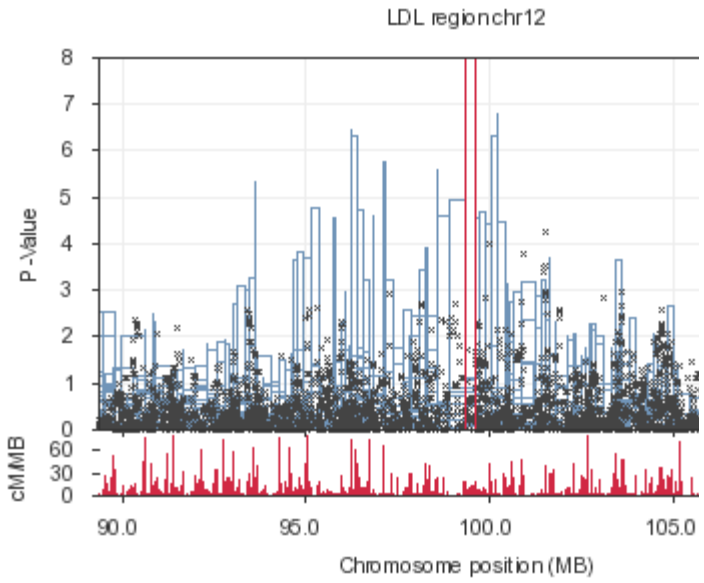
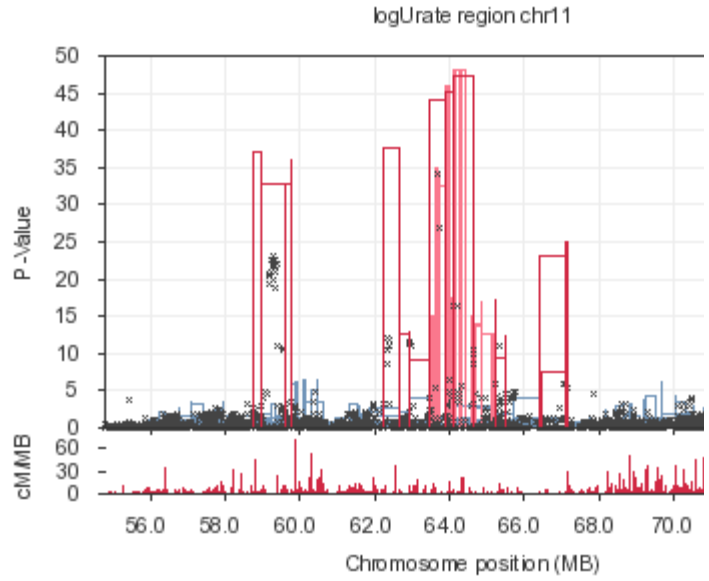


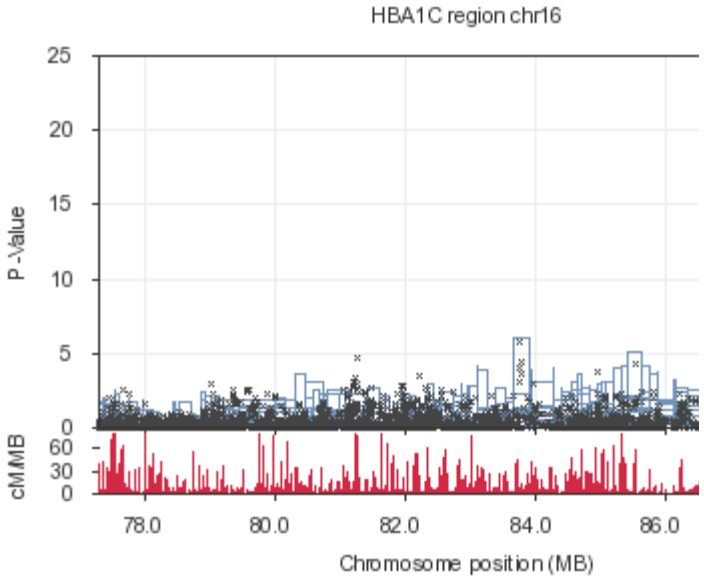
Figure S6. Regions Showing Strong Evidence of Association in European Cohort

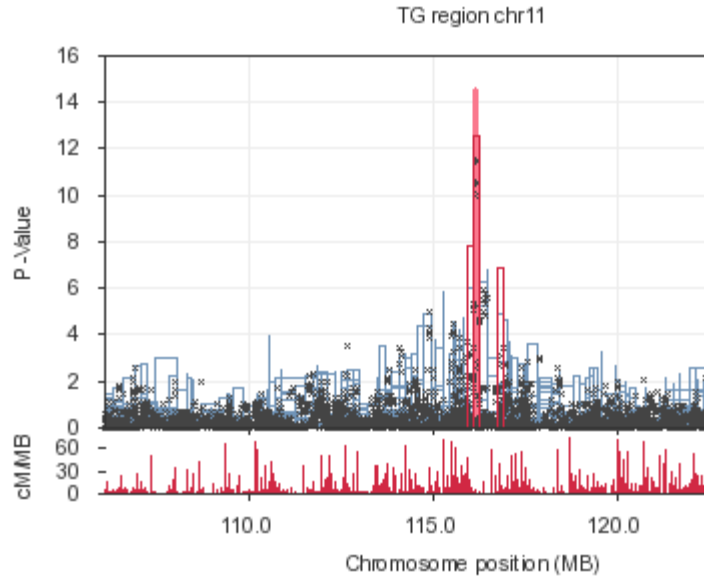
Haplotype boundaries represented as blue bars with height of bar corresponding to allelic *P*-value; genome-wide significant haplotypes shown in red; single marker associations shown as black cross. Lower panel details physical position along chromosome and fine-scale recombination map from HapMap reference.

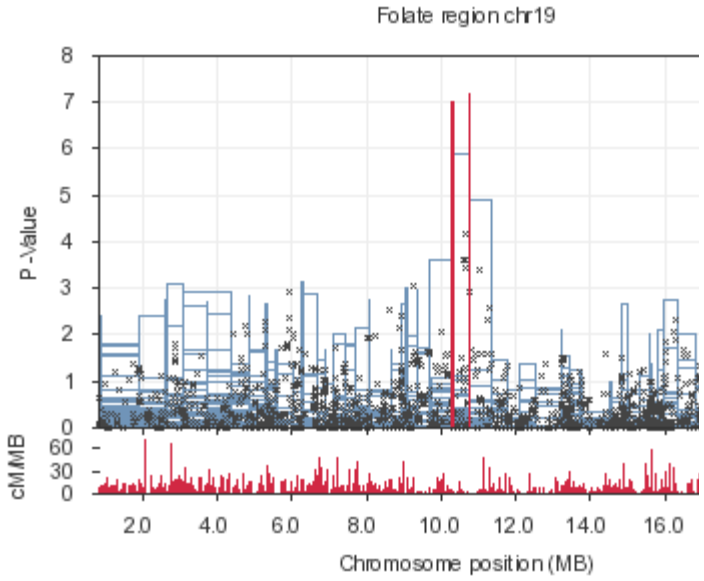

TC region chr6
P -Value
cM/MB
Chromosome position (MB)


LDL region chr12
P -Value
cM/MB
Chromosome position (MB)


logUrate region chr11
P -Value
cM/MB
Chromosome position (MB)


HBA1C region chr16
P -Value
cM/MB
Chromosome position (MB)


TG region chr11
P -Value
cM/MB
Chromosome position (MB)


Folate region chr19
P -Value
cM/MB
Chromosome position (MB)

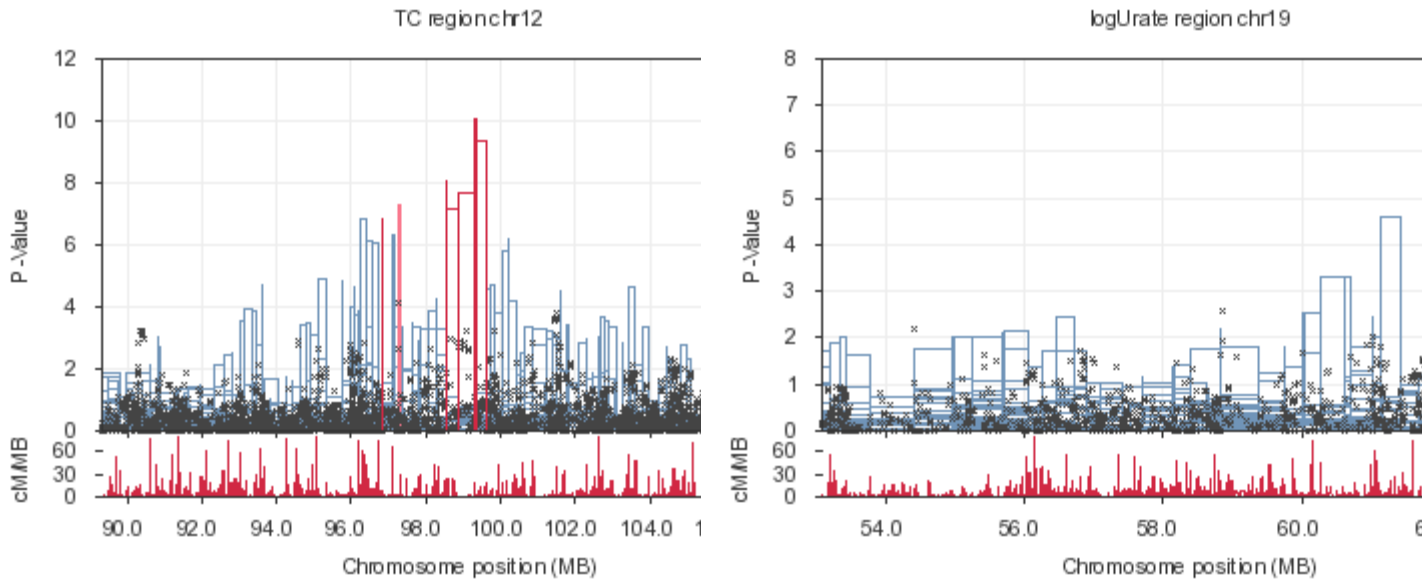


Figure S7. Regions Showing Strong Evidence of Association in Isolate Cohort

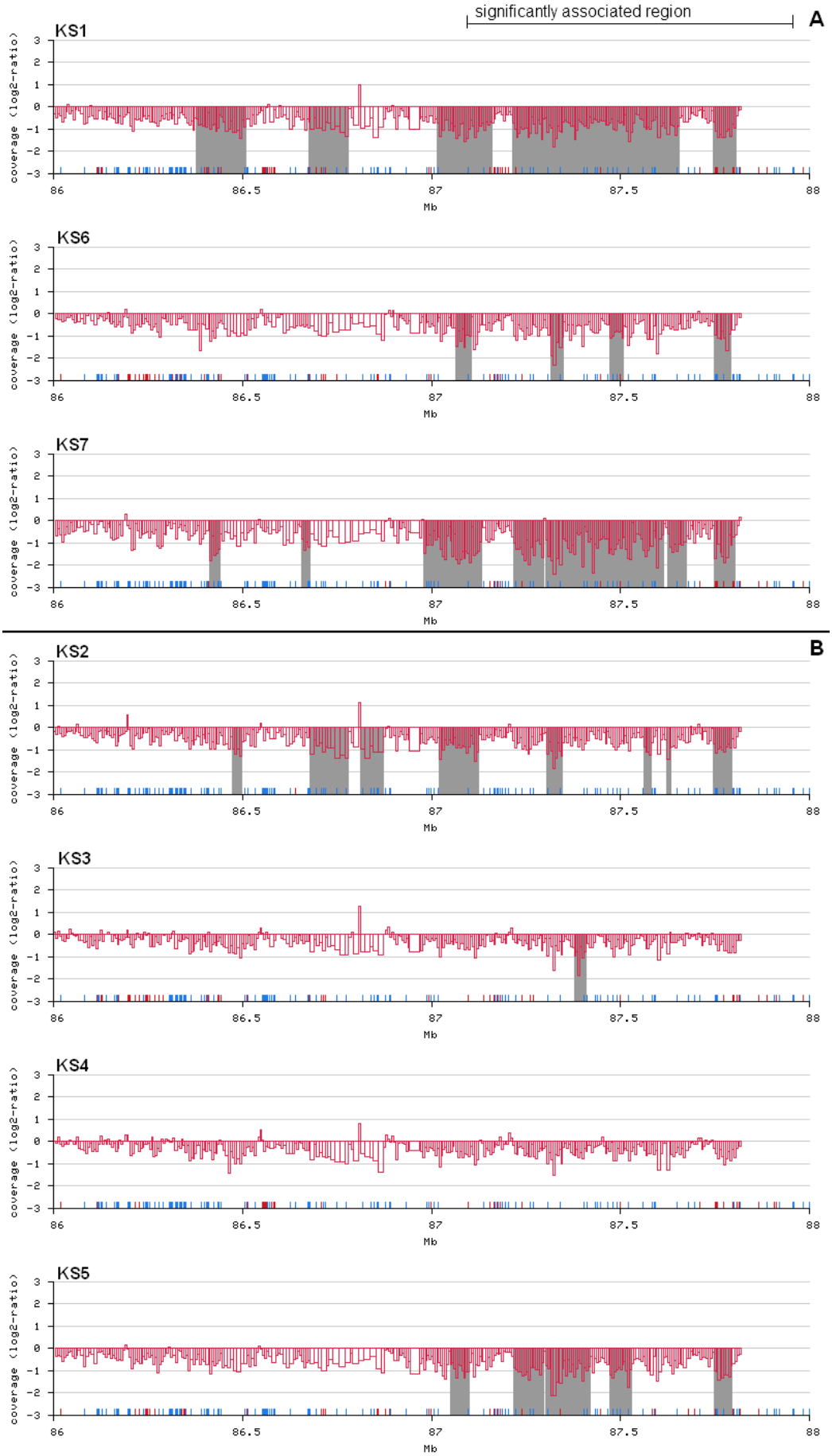
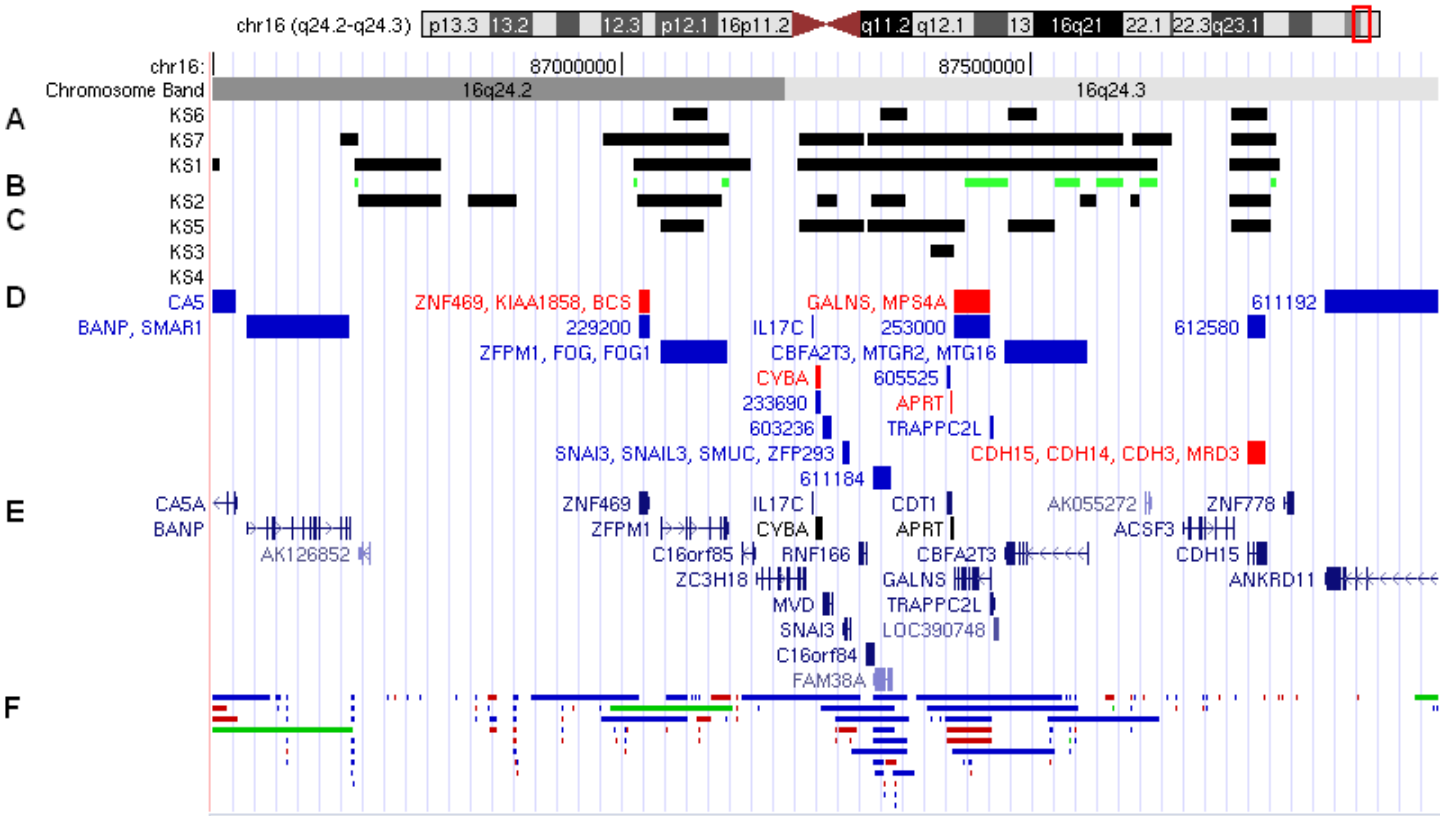


Figure S8. Sequencing Coverage and Detected Copy Loss at Locus of haplotype Association to HBa1c

CNV calling in carriers (**A**) and non-carriers (**B**) of haplotype association on chromosome 16 to Hemoglobin A1c. Red blocks represent window-normalized log2 ratio of coverage; grey regions represent HMM-based heterozygous deletion calls (no other CNVs were present in the region); points along the X-axis correspond to genotyped SNP calls as homozygous (blue) or heterozygous (red).



G

chr	Start	End	Hom. SNPs	Het. SNPs	Overlapping genes
16	87,014,485	87,019,635	9	0	
16	87,123,135	87,131,369	0	0	ZFPM1
16	87,419,927	87,473,134	9	3	CBFA2T3,GALNS,TRAPPC2L,PABPN1L
16	87,530,775	87,561,773	3	0	CBFA2T3
16	87,582,273	87,613,563	9	0	
16	87,634,701	87,655,451	3	0	
16	87,795,956	87,802,189	5	4	

Figure S9. Detected CNV Regions and Carrier-Specific Sites at Locus of Haplotype Association to HBA1c

Genome browser plot detailing structural variants in region of haplotype association to Hemoglobin A1c. Shown are (**A**) CNVs present in carriers; (**B**, green) CNV regions exclusively in carriers; (**C**) CNVs present in non-carriers; (**D**) OMIM genes, (**E**) UCSC known genes; and (**F**) known CNVs from the Database of Genomic Variants. Table in (**G**) details carrier exclusive regions, observed homozygosity in Affymetrix 500k SNPs and overlapping exons.